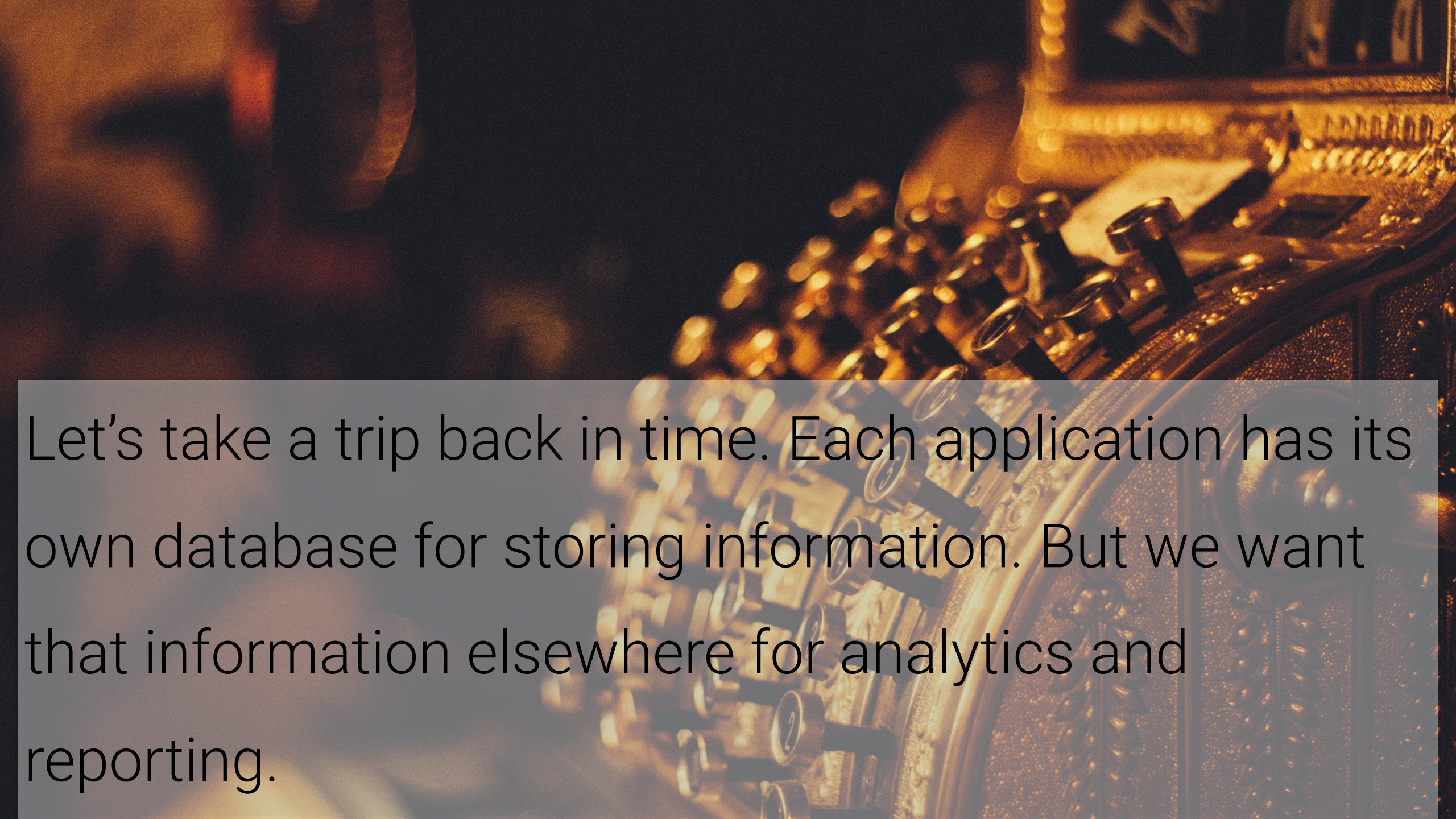
https://speakerdeck.com/rmoff/

Look Ma, no Code! Building Streaming Data Pipelines with Apache Kafka

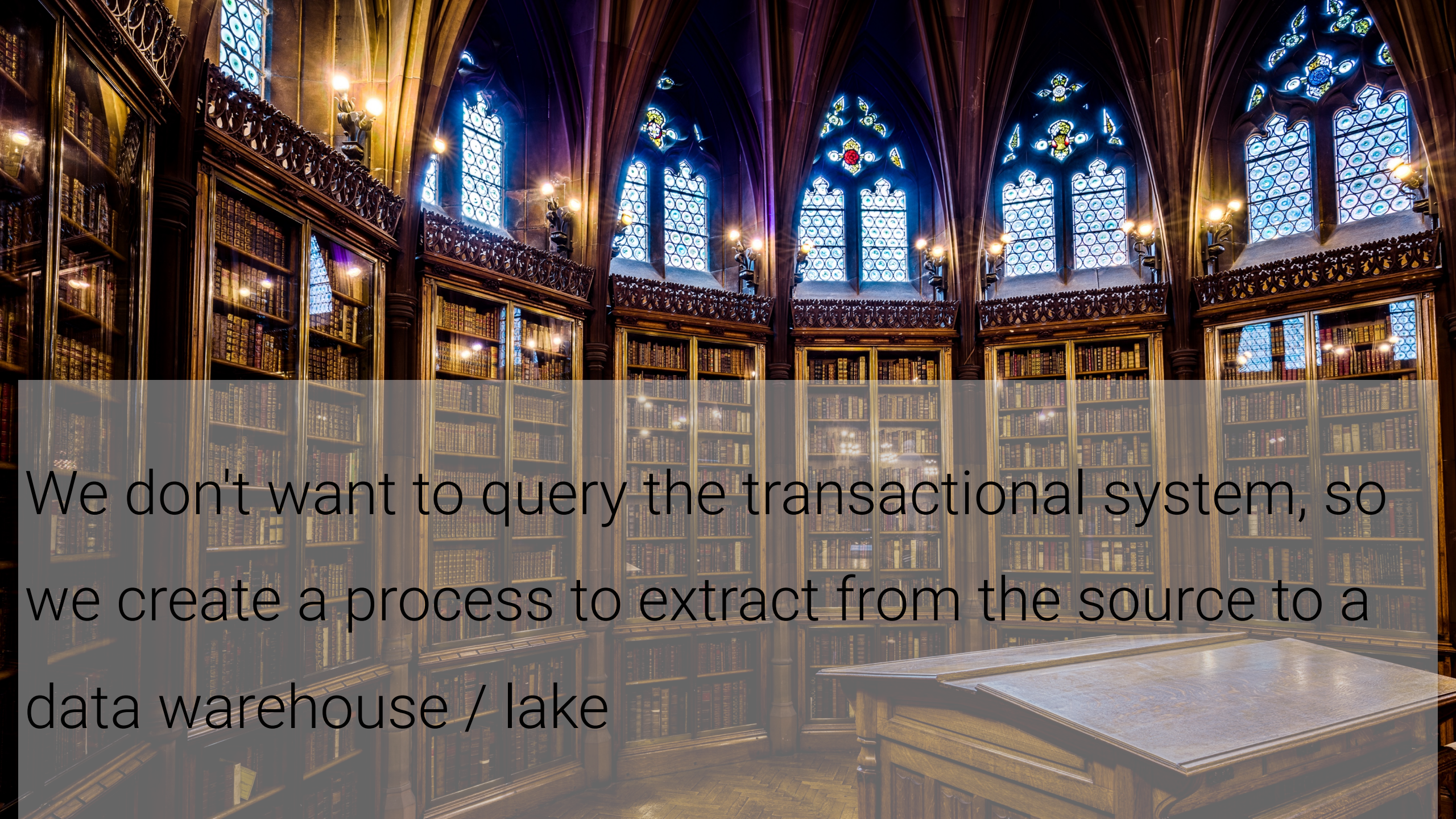
Kafka Summit London

24 Apr 2018 / Robin Moffatt

@rmoff robin@confluent.io



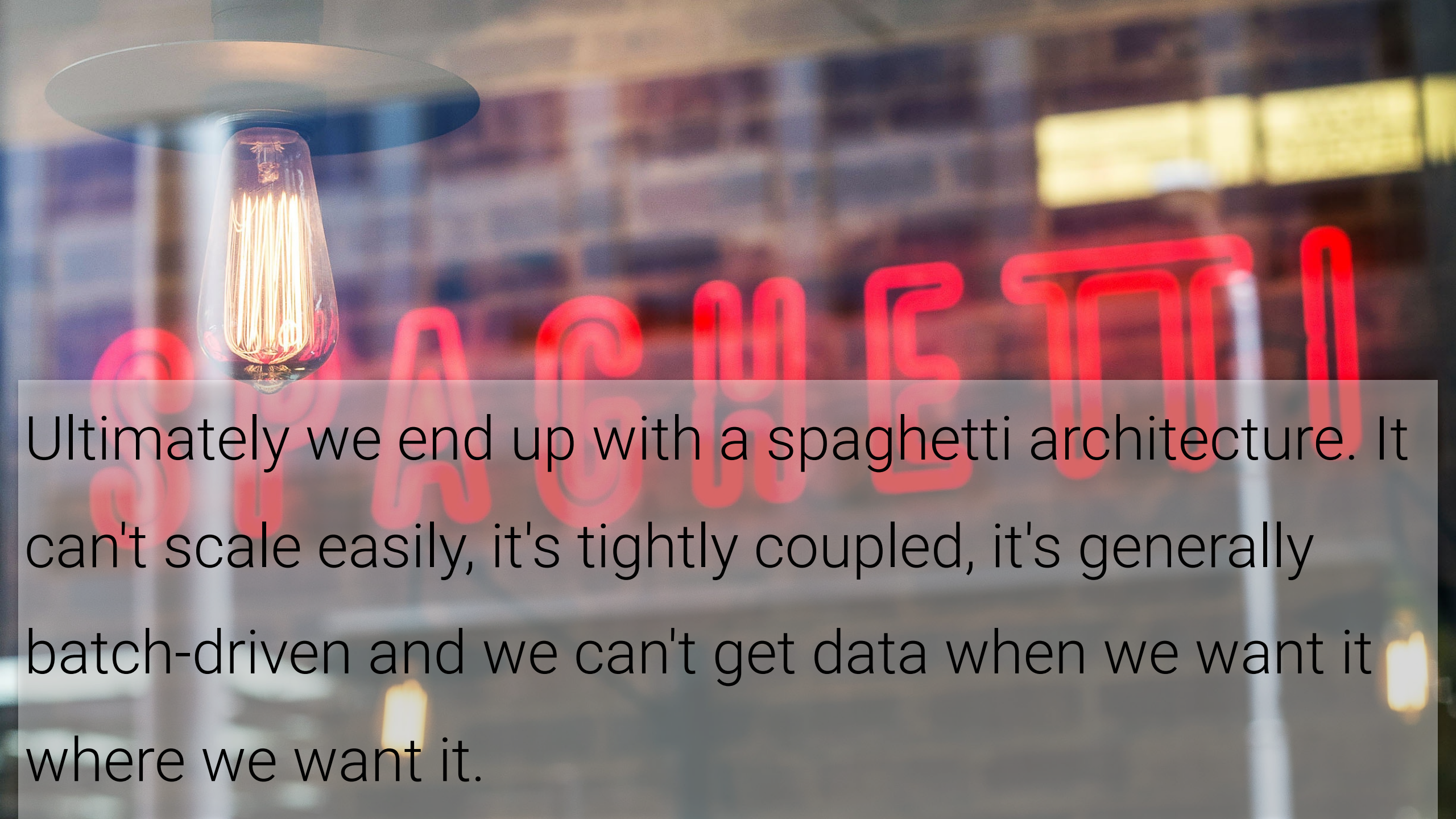
Let’s take a trip back in time. Each application has its own database for storing information. But we want that information elsewhere for analytics and reporting.



We don't want to query the transactional system, so we create a process to extract from the source to a data warehouse / lake

We want to unify data from multiple systems, so create conformed dimensions and batch processes to federate our data. This is all batch driven, so latency is built in by design.

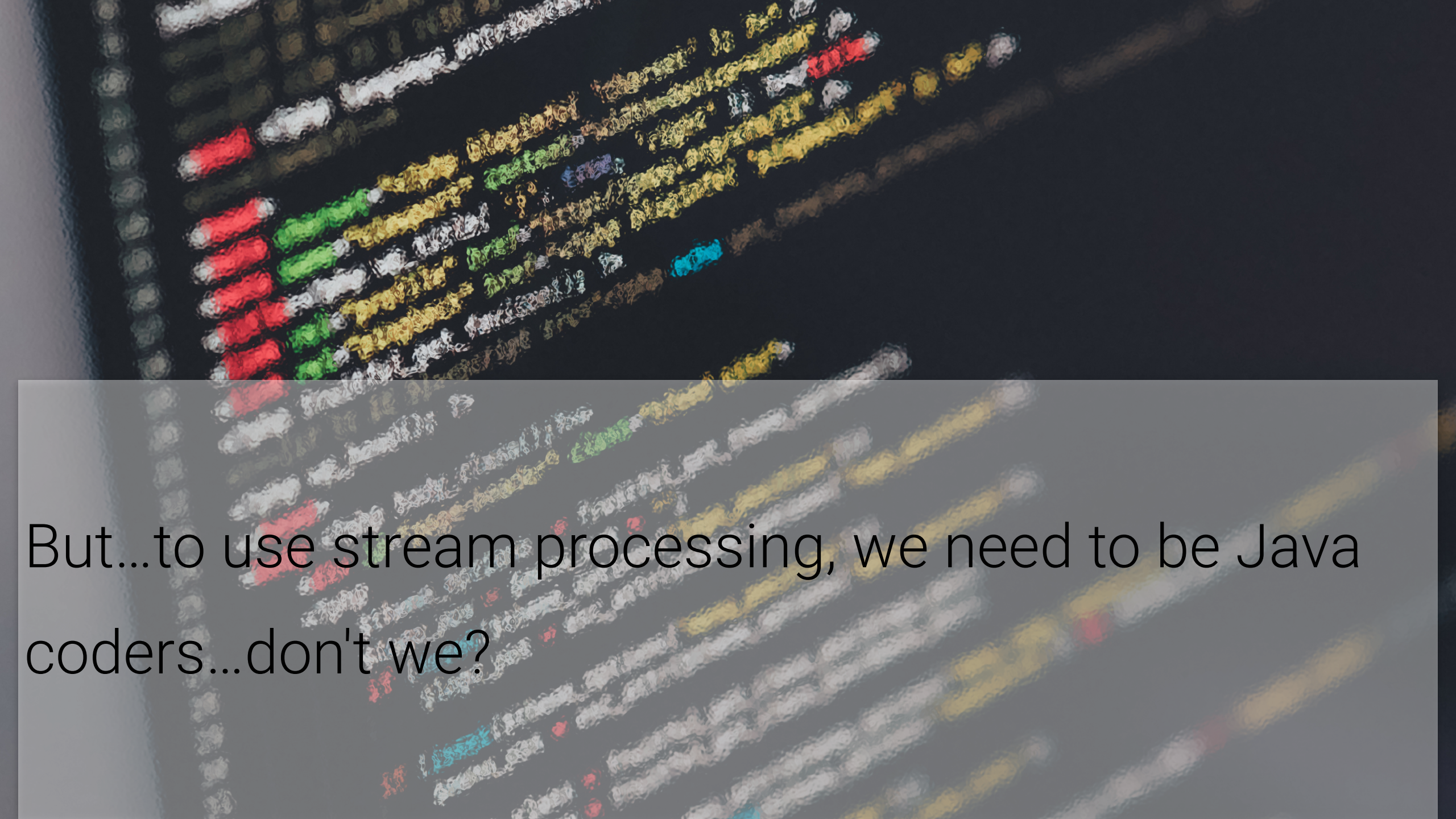
As well as our data warehouse, we want to use our transactional data to populate search replicas, Graph databases, noSQL stores...all introducing more point-to-point dependencies in our system



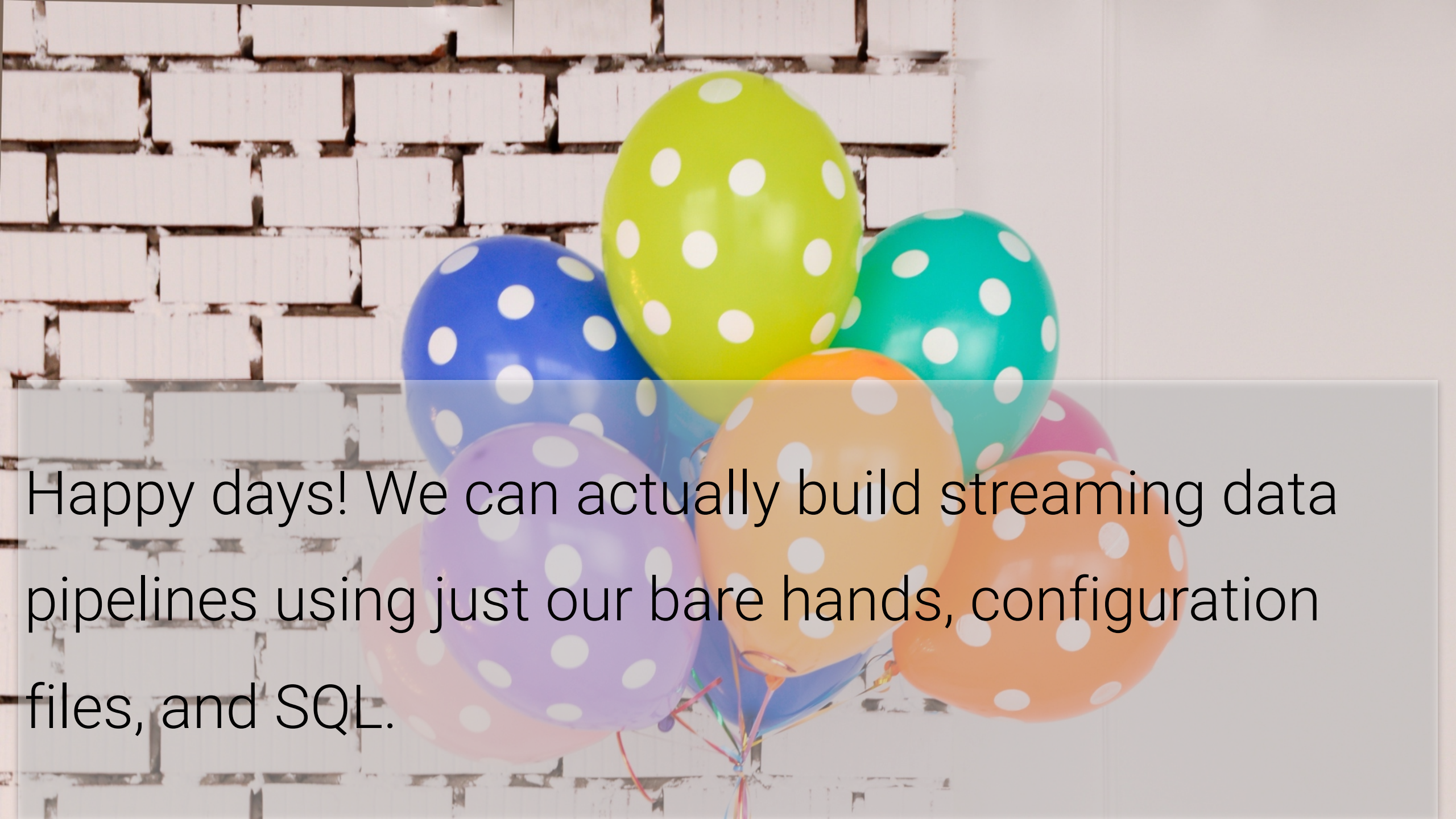
Ultimately we end up with a spaghetti architecture. It can't scale easily, it's tightly coupled, it's generally batch-driven and we can't get data when we want it where we want it.

But...there's hope!

Apache Kafka, a distributed streaming platform, enables us to decouple all our applications creating data from those utilising it. We can create low-latency streams of data, transformed as necessary.

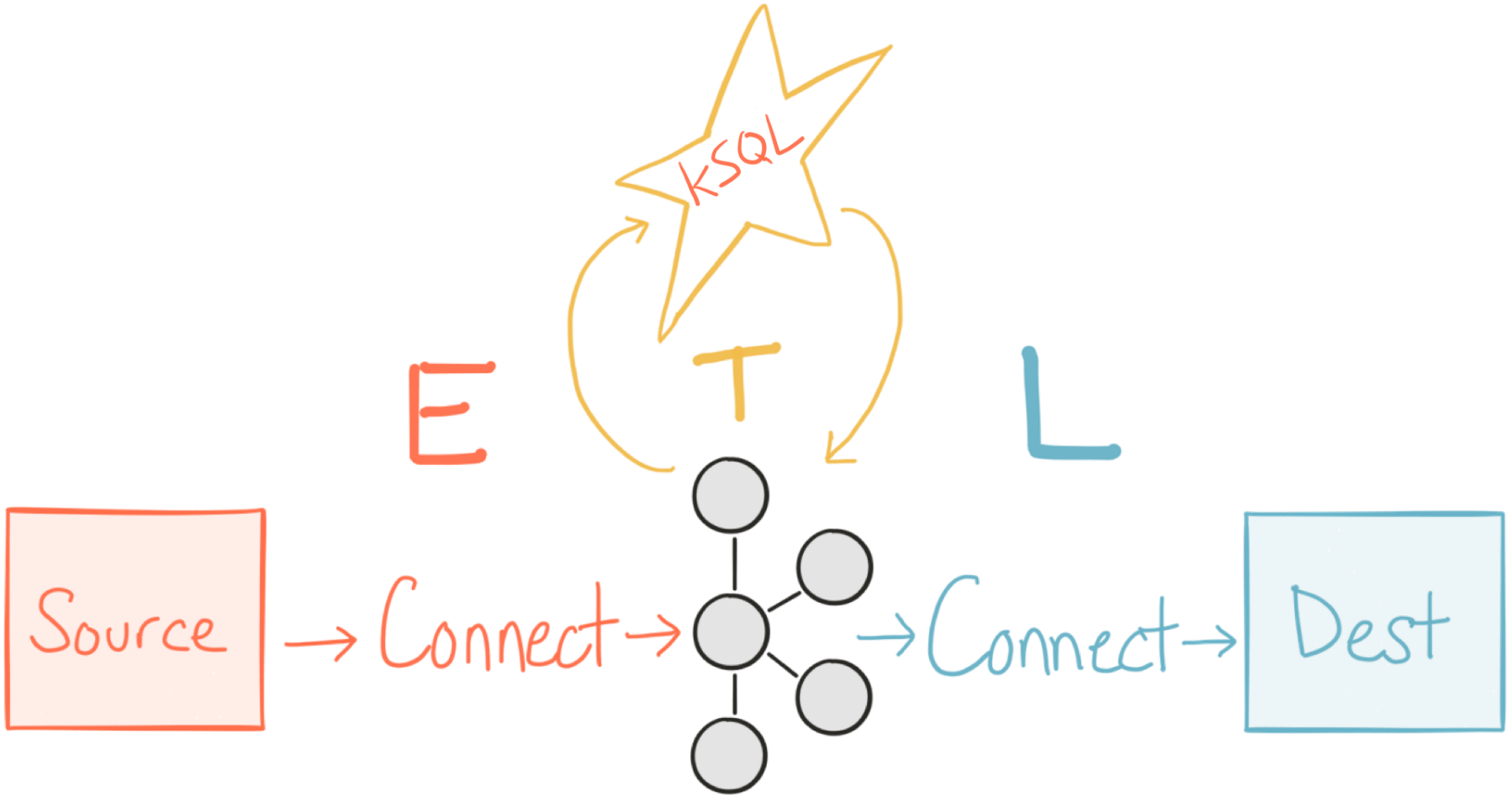


But...to use stream processing, we need to be Java coders...don't we?



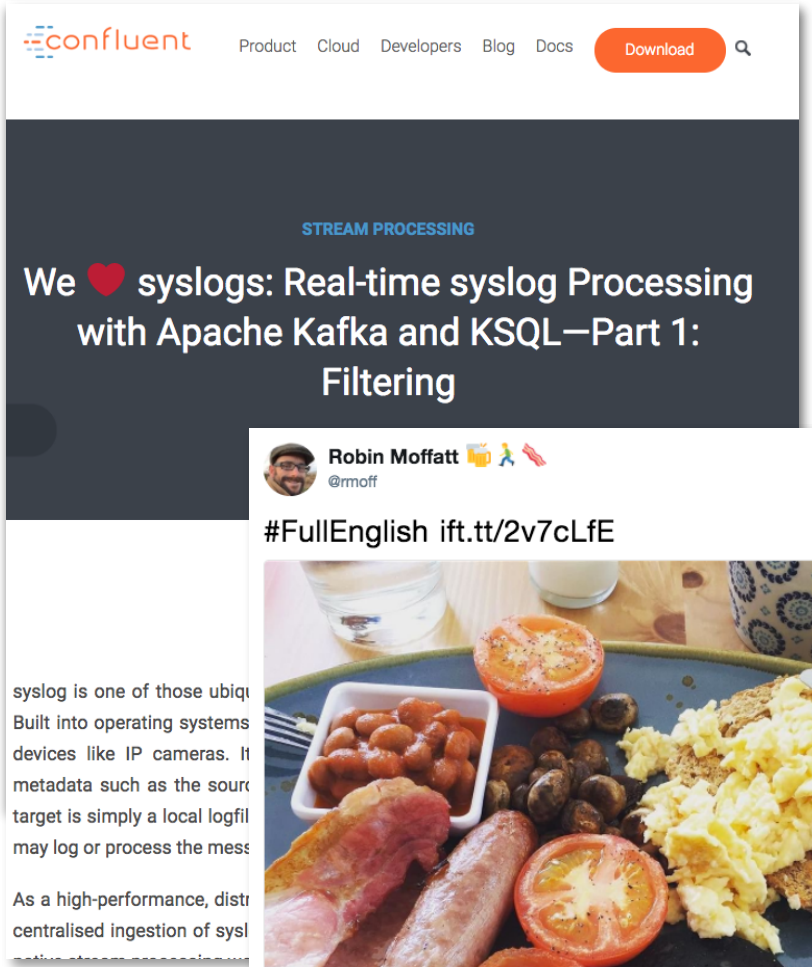
Happy days! We can actually build streaming data pipelines using just our bare hands, configuration files, and SQL.

Streaming ETL, with Apache Kafka and Confluent Platform



$ whoami

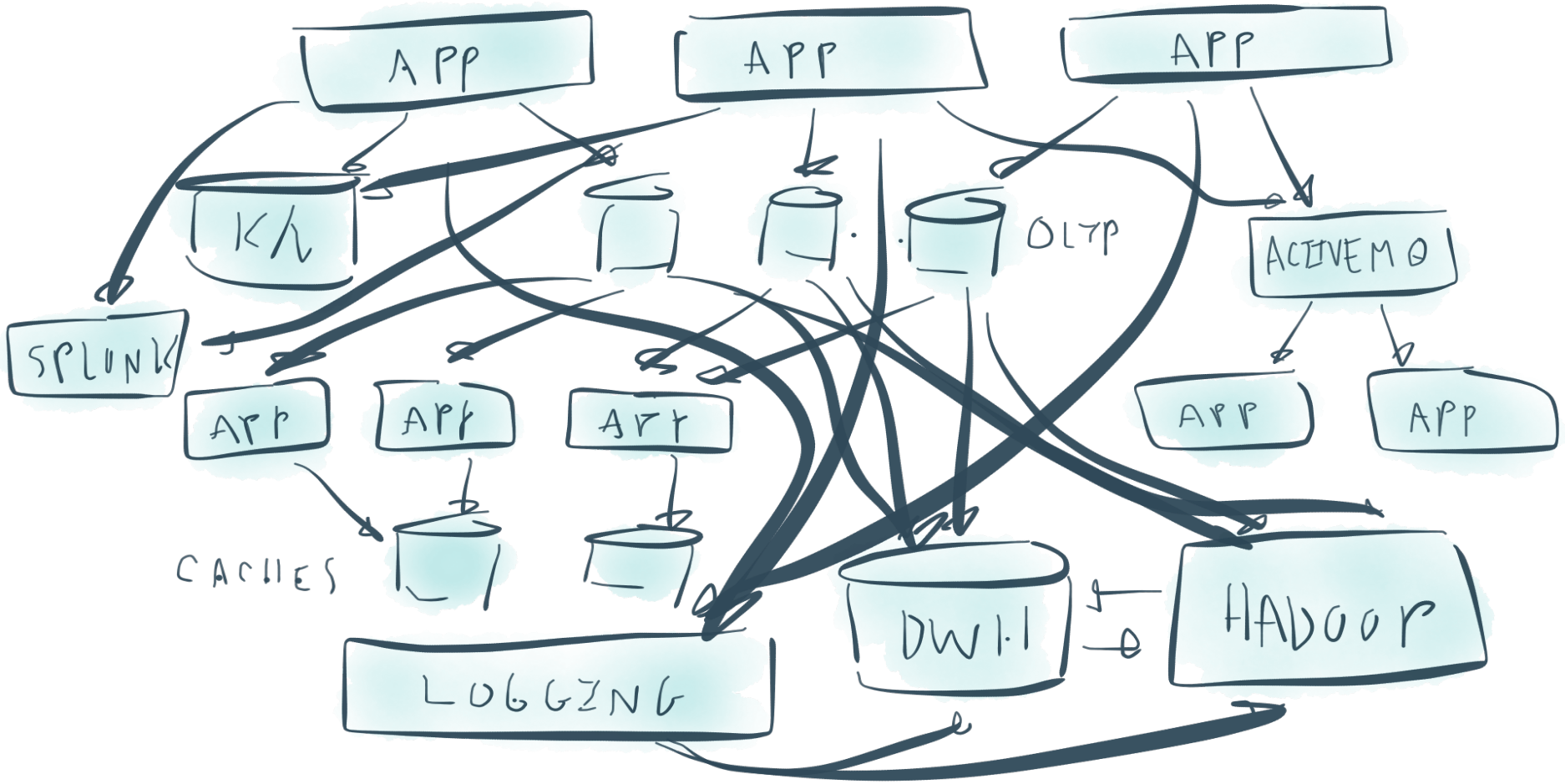
- Partner Technology Evangelist @ Confluent
- Working in data & analytics since 2001
- Oracle ACE Director & Dev Champion
- Blogging : http://rmoff.net & https://www.confluent.io/blog/author/robin/
- Twitter: @rmoff
  - Geek stuff
  - Beer & Fried Breakfasts



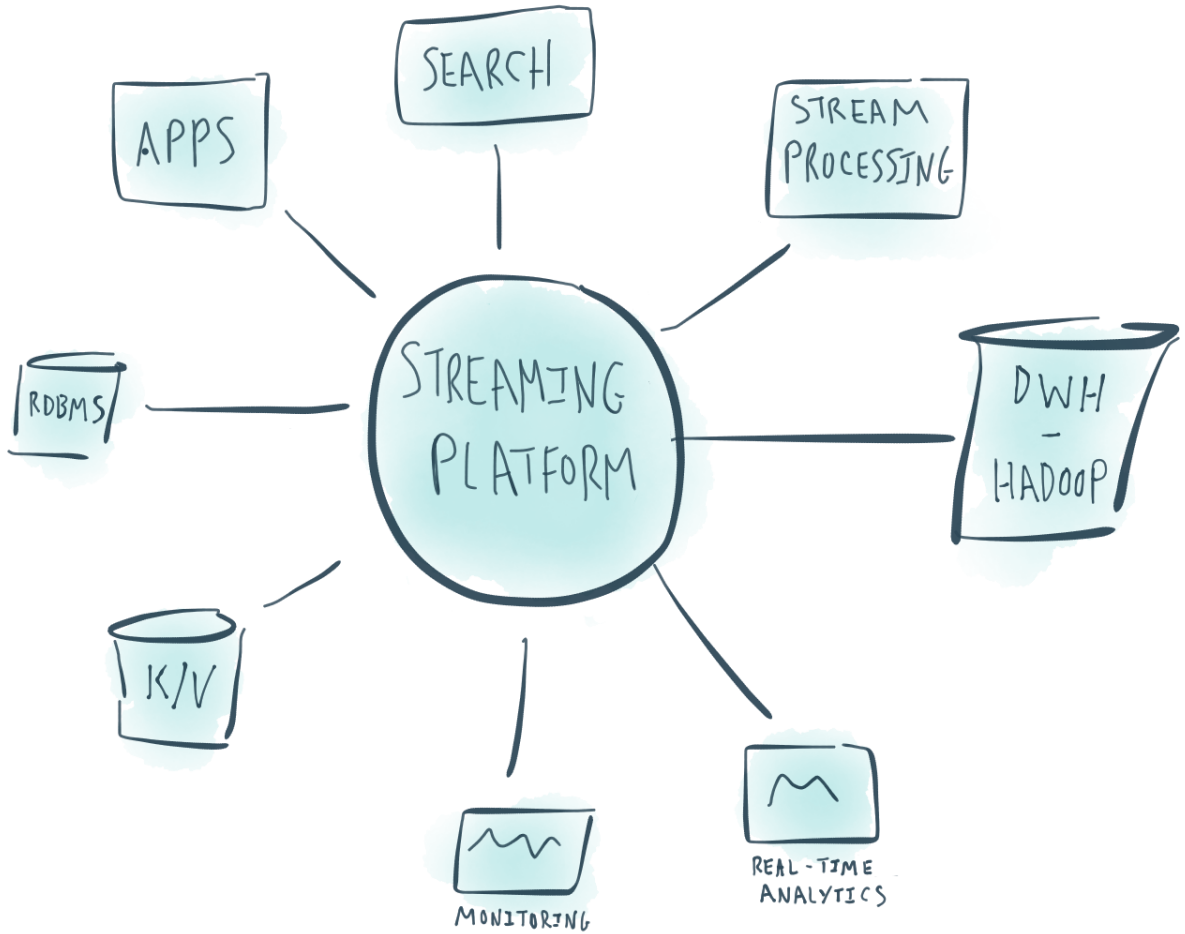
https://speakerdeck.com/rmoff/

confluent

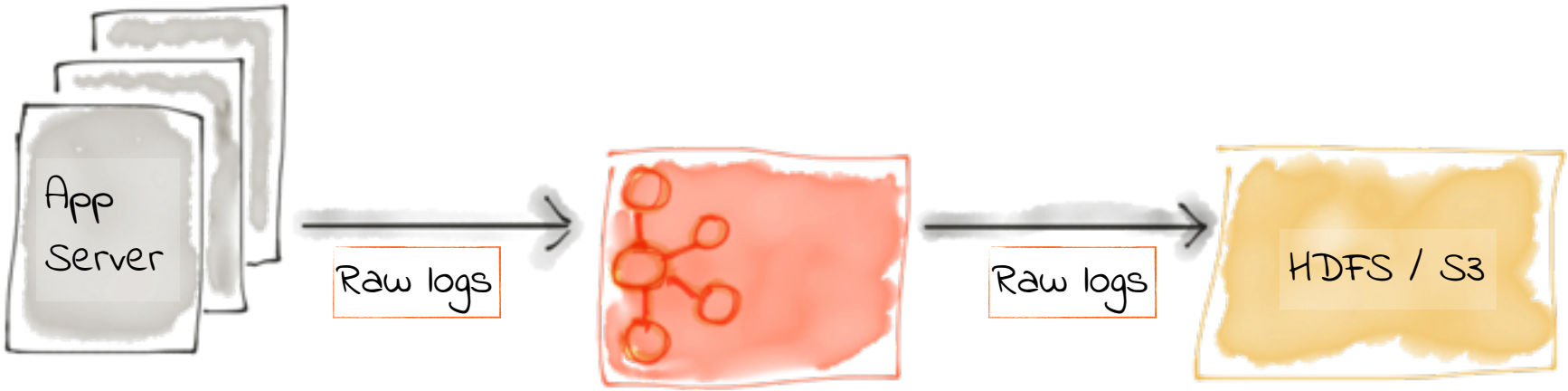
A GIANT MESS


APP
APP
APP
K/L
OLTP
ACTIVEMQ
SPLUNK
APP
APP
APP
APP
APP
CACHES
DW1.1
HADOOP
LOGGING

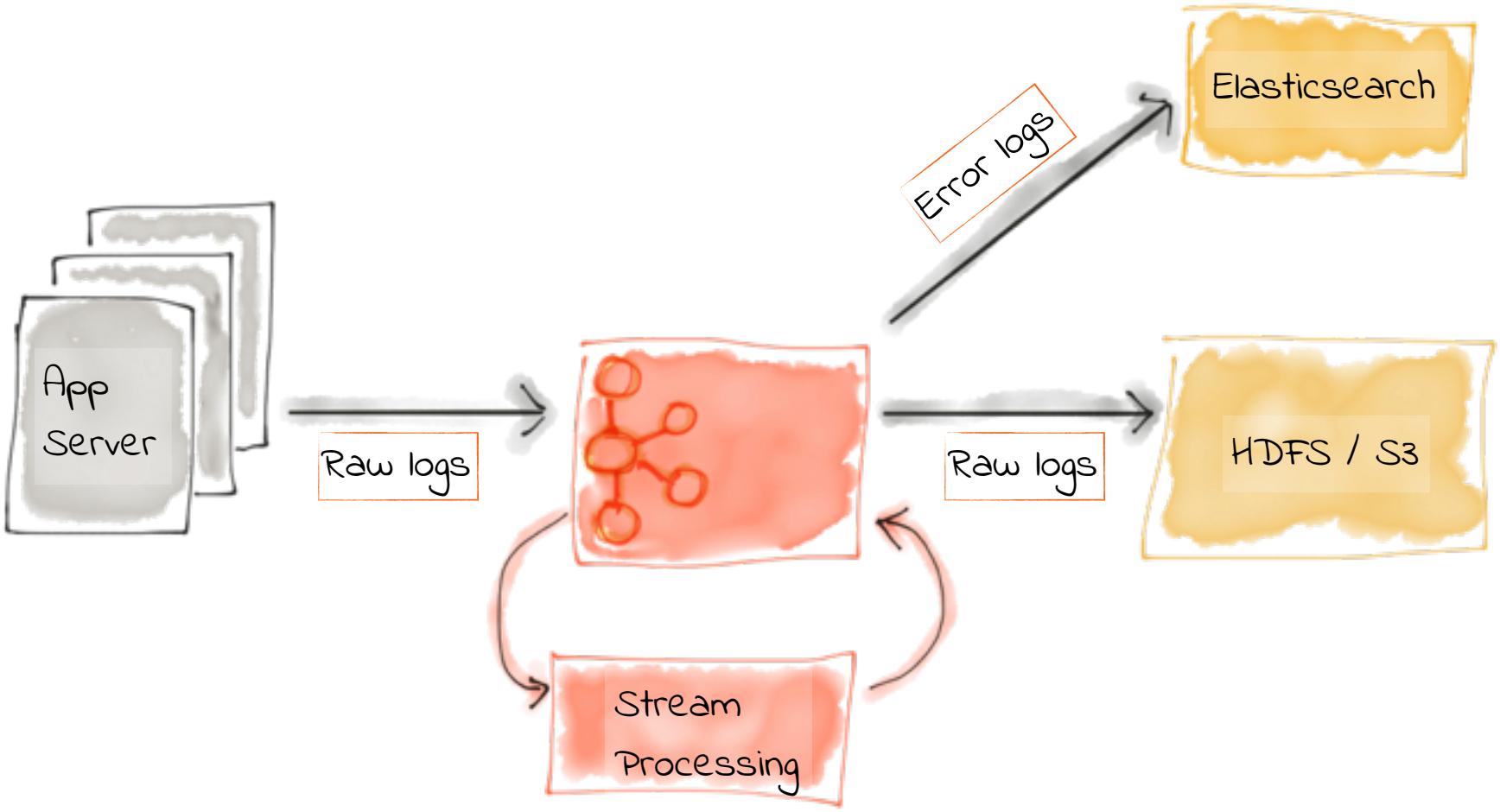
STREAMING PLATFORM



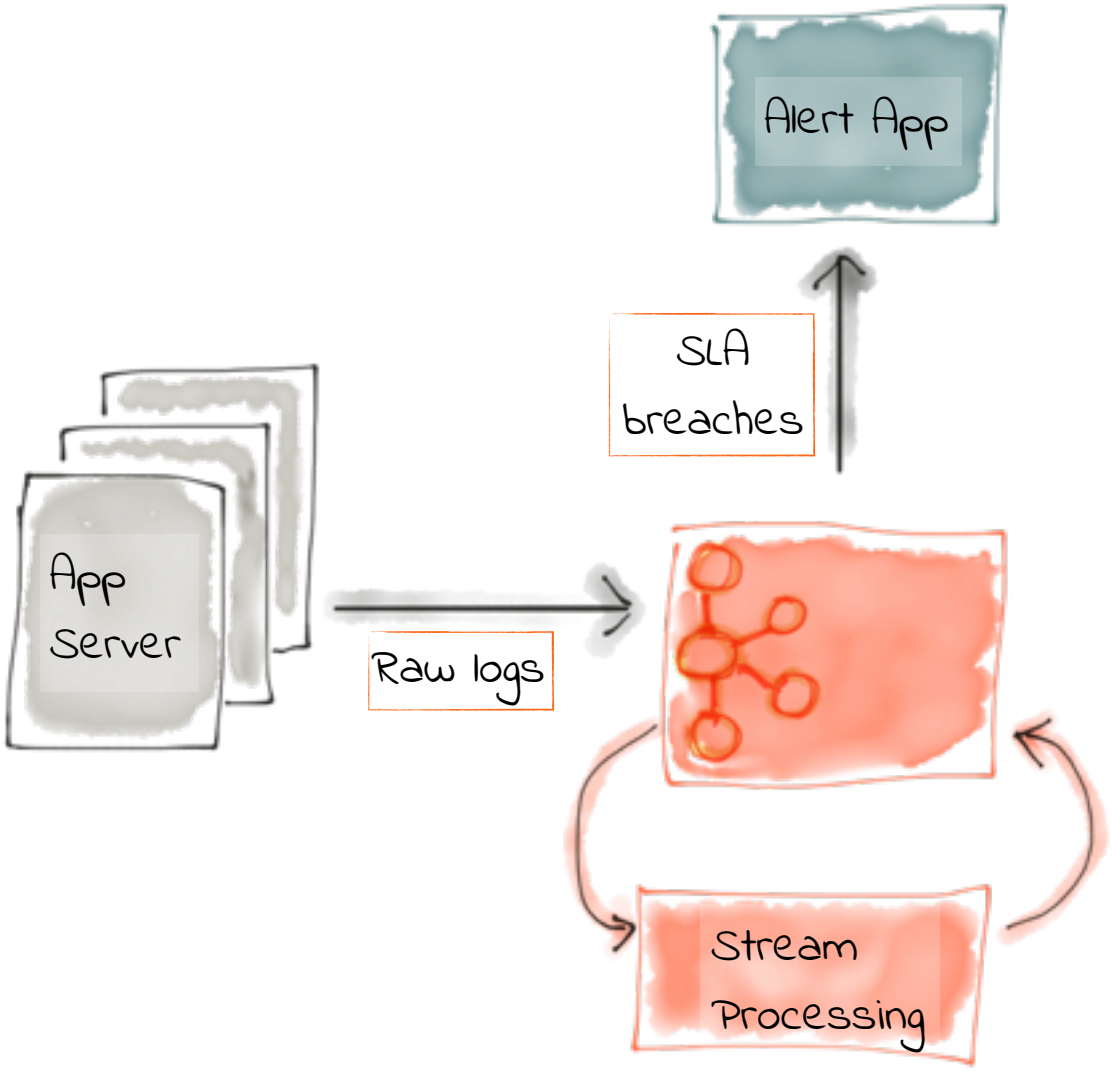
Build Pipelines



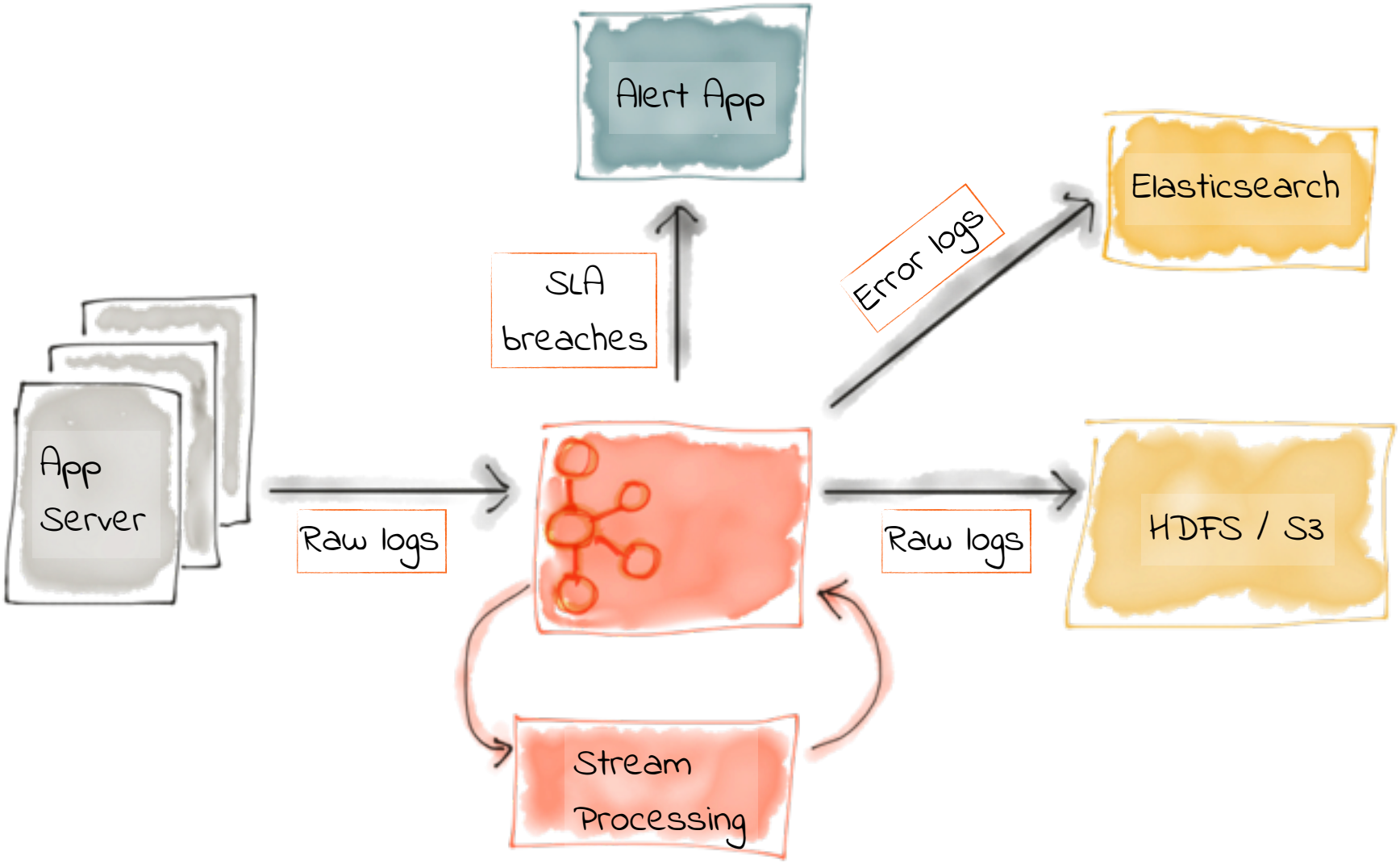
Build Complex Pipelines



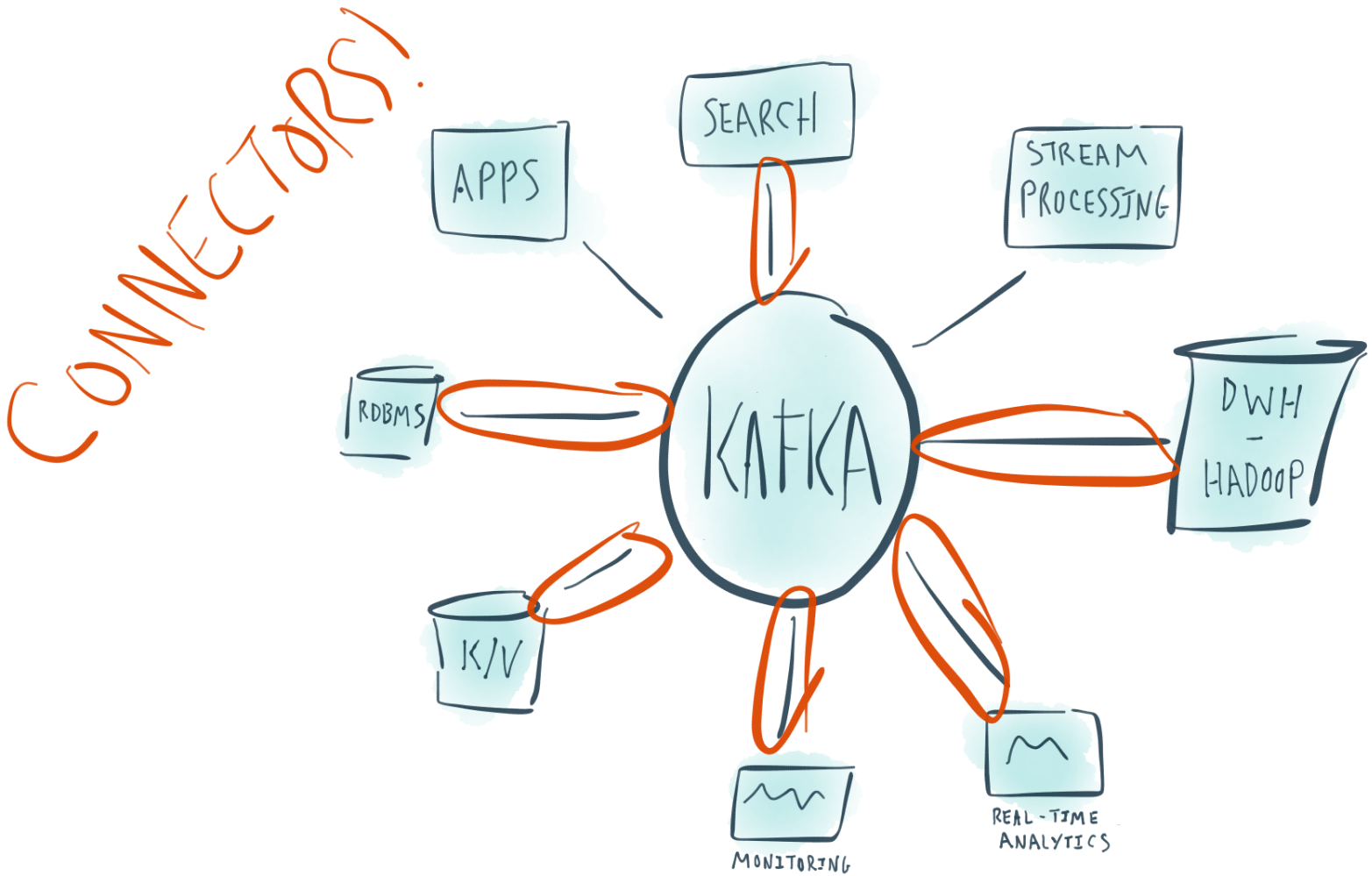
Build Applications



Build Applications + Pipelines



STREAMING PLATFORM



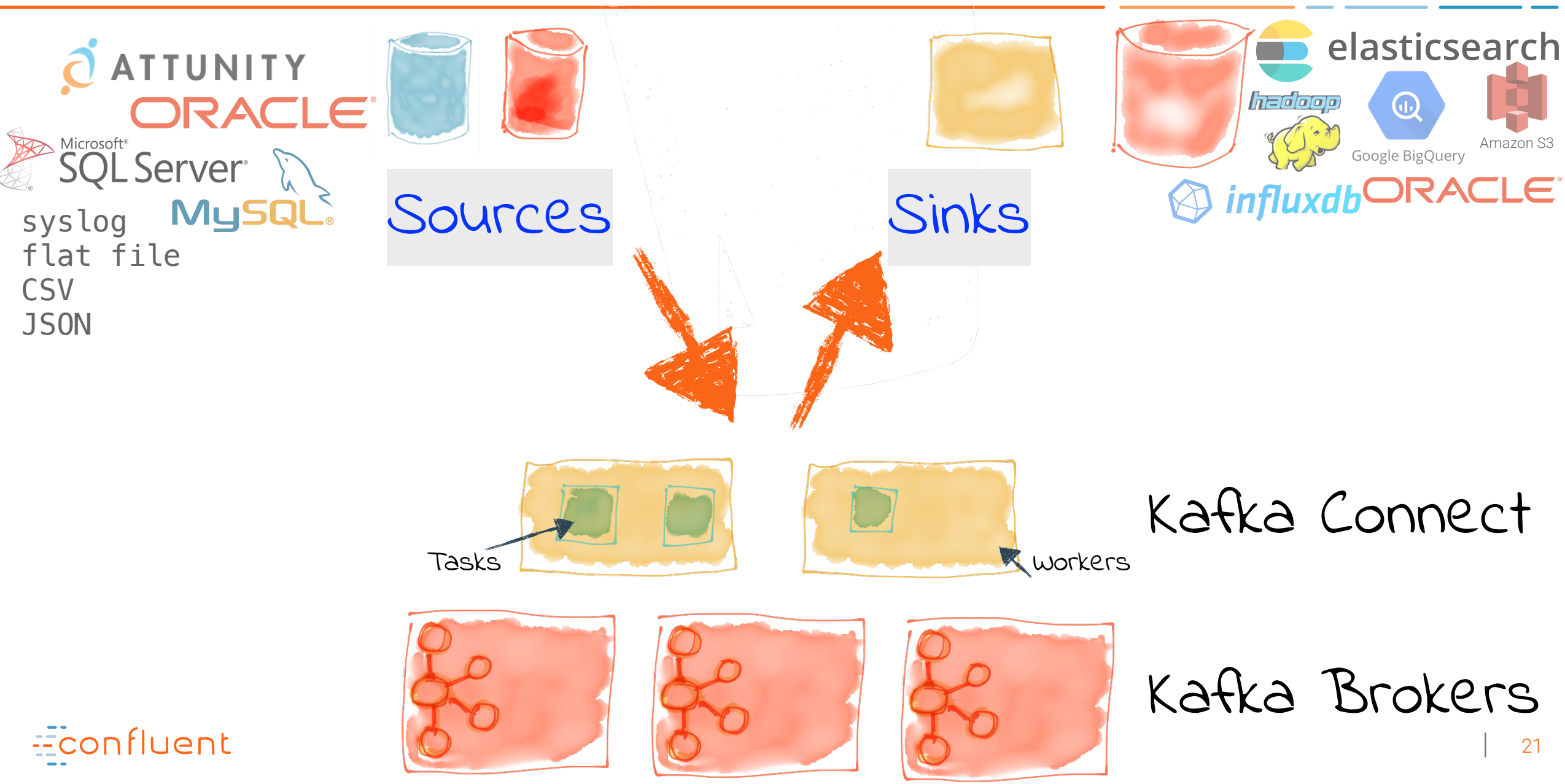
The Connect API of Apache Kafka®

Reliable and scalable integration of Kafka with other systems – no coding required.

- ✓ Centralized management and configuration
- ✓ Support for hundreds of technologies including RDBMS, Elasticsearch, HDFS, S3
- ✓ Supports CDC ingest of events from RDBMS
- ✓ Preserves data schema
- ✓ Fault tolerant and automatically load balanced
- ✓ Extensible API
- ✓ Single Message Transforms
- ✓ Part of Apache Kafka, included in Confluent Open Source

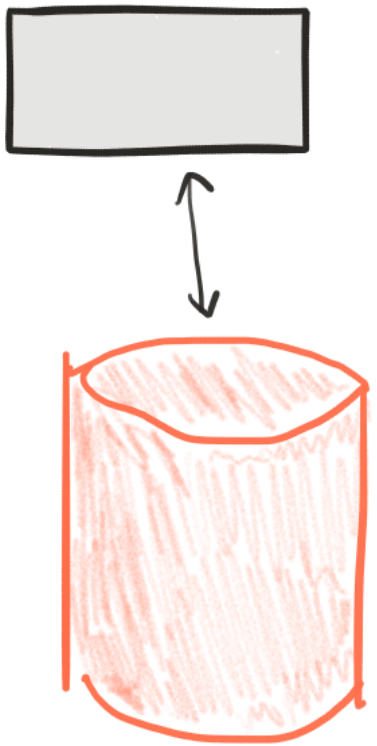
```
{
"connector.class": "io.confluent.connect.jdbc.JdbcSourceConnector",
"connection.url": "jdbc:mysql://localhost:3306/demo?user=rmoff&password=foo",
"table.whitelist": "sales,orders,customers"
}
```

Kafka Connect



Streaming Application Data to Kafka

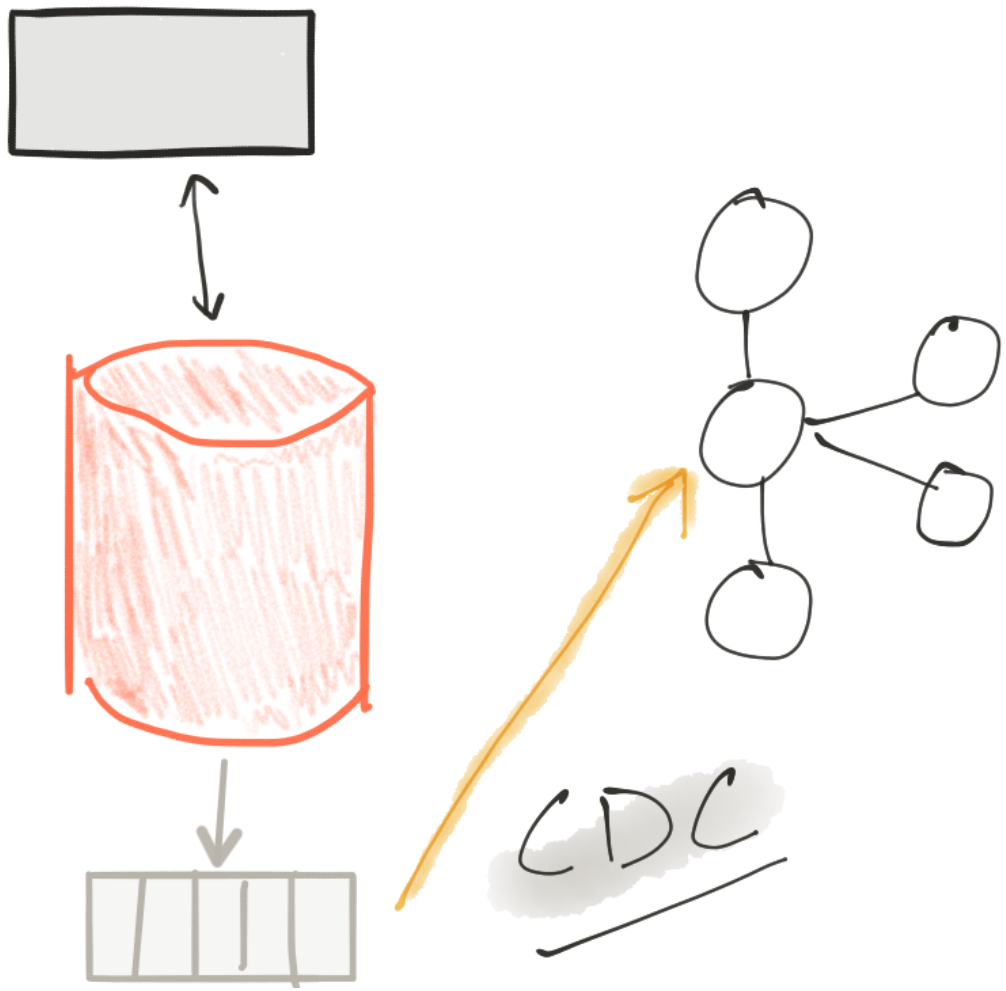
- Modifying applications is not always possible or desirable
  - And what if the data gets changed within the database or by other apps?
- Query-based CDC with Kafka Connect's JDBC source is one option for extracting data
  - Polls the source DB --> Latency & Load implications
  - Requires key/ts in the schema
  - Can't capture deletes



```
"name": "jdbc_source_mysql_01",
"config": {
  "connector.class": "io.confluent.connect.jdbc.JdbcSourceConnector",
  "connection.url": "jdbc:mysql://localhost:3306/sakila?user=rmoff&passwor
```

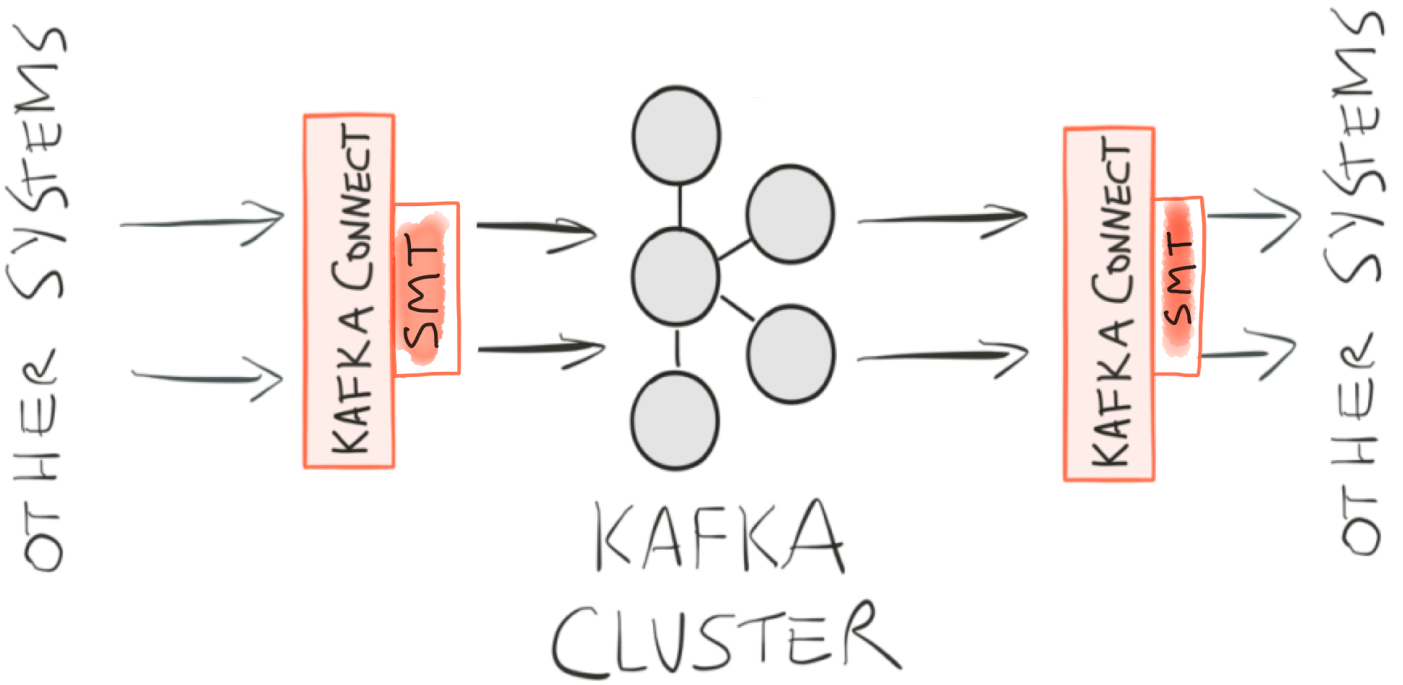
Liberate Application Data into Kafka with log-based CDC

- Relational databases use transaction logs to ensure Durability of data
- Log-based Change-Data-Capture (CDC) mines the log to get raw events from the database
- CDC tools that integrate with Kafka Connect include:
  - Debezium
  - GoldenGate
  - Attunity
  - + more



Single Message Transform (SMT) -- Extract, TRANSFORM, Load…

- Modify events *before storing* in Kafka:
  - Mask/drop sensitive information
  - Set partitioning key
  - Store lineage
- Modify events *going out* of Kafka:
  - Route high priority events to faster data stores
  - Direct events to different Elasticsearch indexes
  - Cast data types to match destination



But I need to

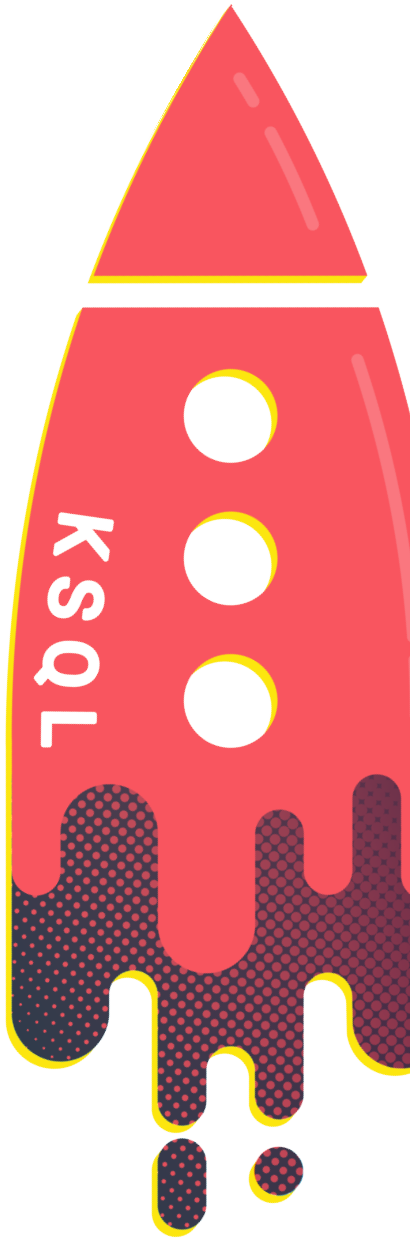
join...aggregate...filter...

KSQL

is a

Declarative

Stream Processing Language



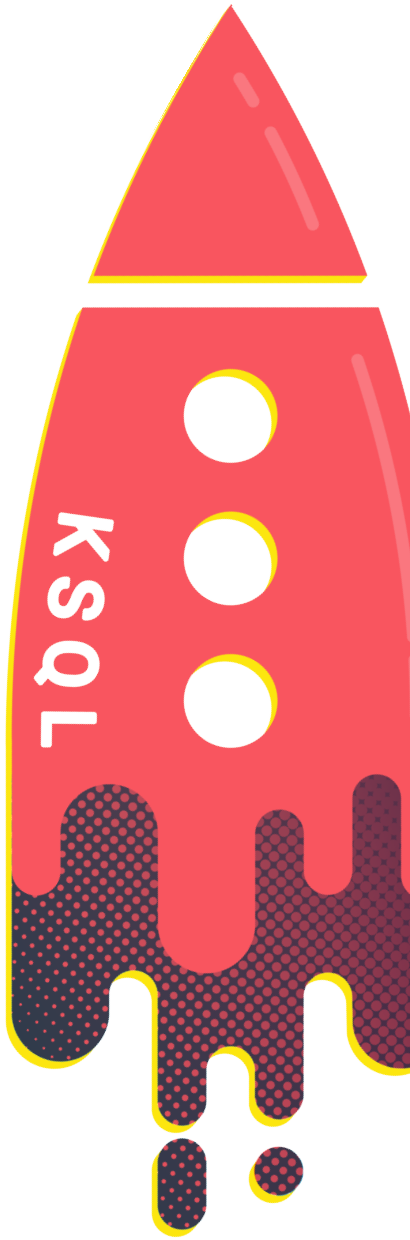
KSQL

is the

Streaming SQL Engine

for

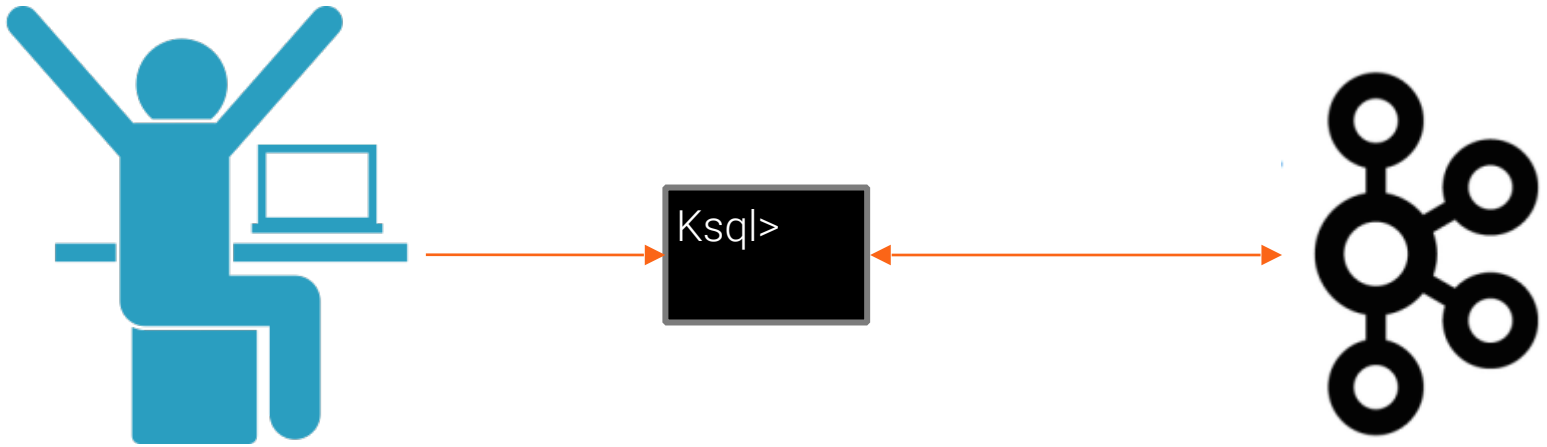
Apache Kafka



confluent

KSQL: a Streaming SQL Engine for Apache Kafka™ from Confluent

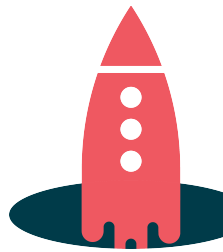
- Enables stream processing with zero coding required
- The simplest way to process streams of data in real-time
- Powered by Kafka: scalable, distributed, battle-tested
- All you need is Kafka–No complex deployments of bespoke systems for stream processing



KSQL for Streaming ETL

Joining, filtering, and aggregating streams of event data

```
CREATE STREAM vip_actions AS
  SELECT userid, page, action
  FROM clickstream c
  LEFT JOIN users u
    ON c.userid = u.user_id
  WHERE u.level = 'Platinum';
```

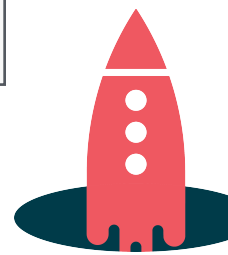


confluent

KSQL for Anomaly Detection

Identifying patterns or anomalies in real-time data, surfaced in milliseconds

```
CREATE TABLE possible_fraud AS
  SELECT card_number, count(*)
    FROM authorization_attempts
    WINDOW TUMBLING (SIZE 5 SECONDS)
    GROUP BY card_number
    HAVING count(*) > 3;
```

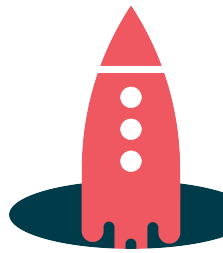


confluent

KSQL for Real-Time Monitoring

- Log data monitoring, tracking and alerting
- syslog data
- Sensor / IoT data

```
CREATE STREAM SYSLOG_INVALID_USERS AS
SELECT HOST, MESSAGE
FROM SYSLOG
WHERE MESSAGE LIKE '%Invalid user%';
```



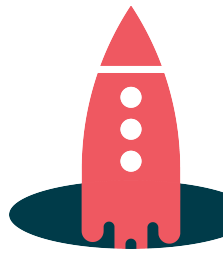
confluent

http://cnfl.io/syslogs-filtering / http://cnfl.io/syslog-alerting

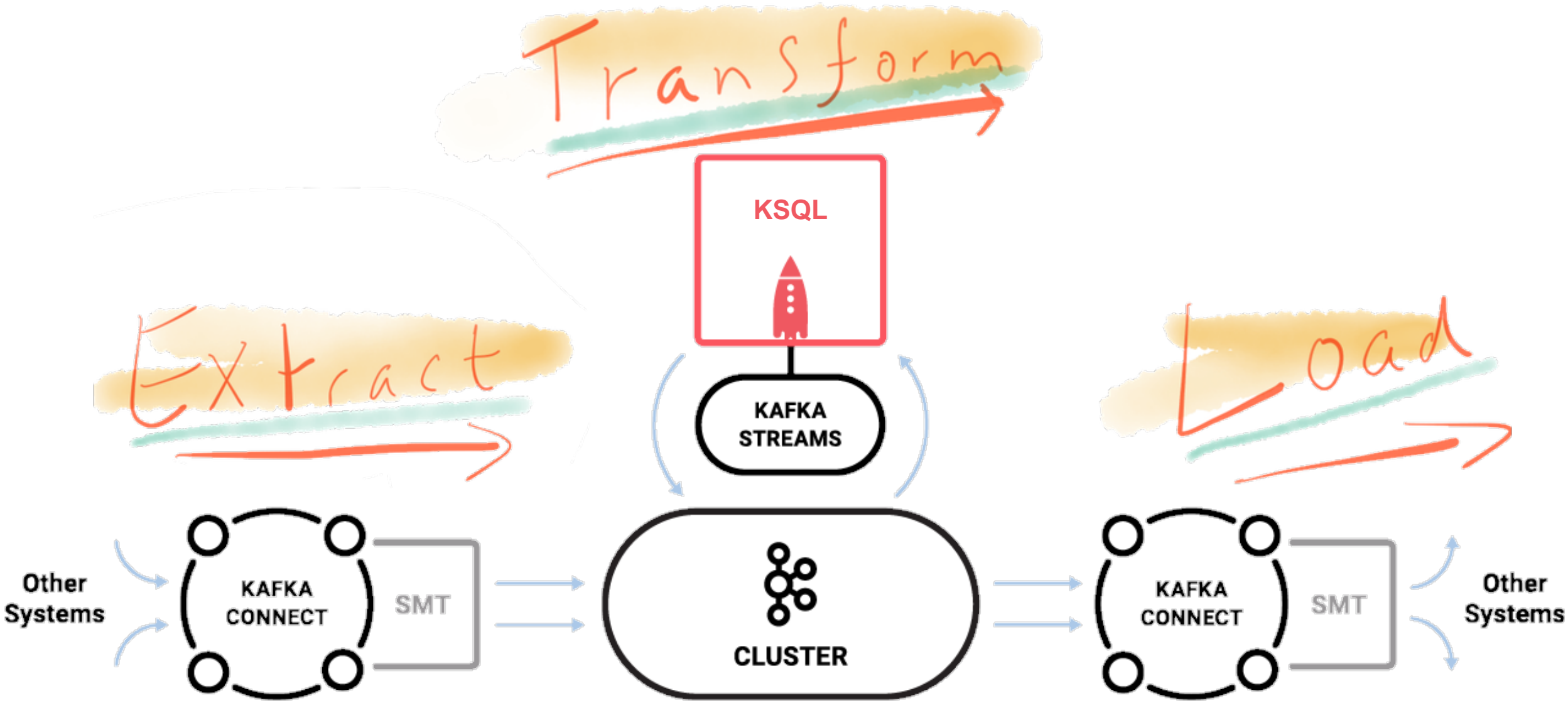
KSQL for Data Transformation

Make simple derivations of existing topics from the command line

```
CREATE STREAM views_by_userid
  WITH (PARTITIONS=6, REPLICAS=5,
        VALUE_FORMAT='AVRO',
        TIMESTAMP='view_time') AS
  SELECT * FROM clickstream
  PARTITION BY user_id;
```



Streaming ETL, powered by Apache Kafka and Confluent Platform



Demo Time!

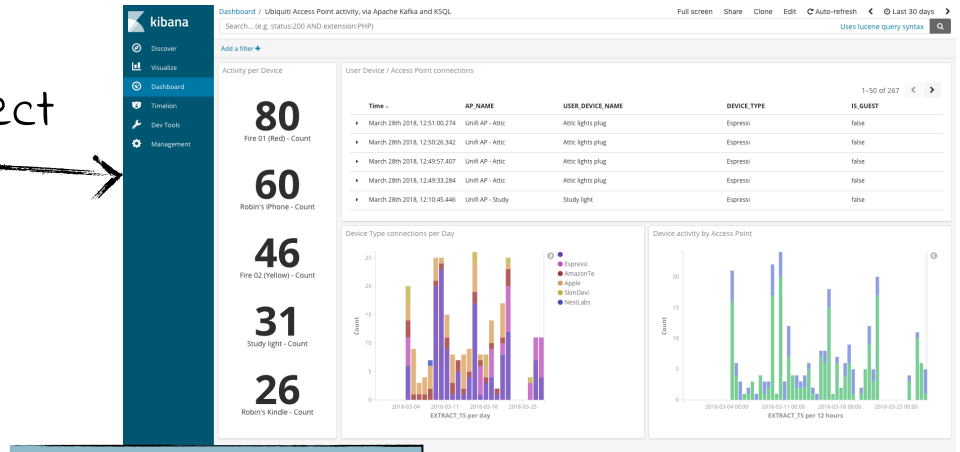
ratings stream

MySQL

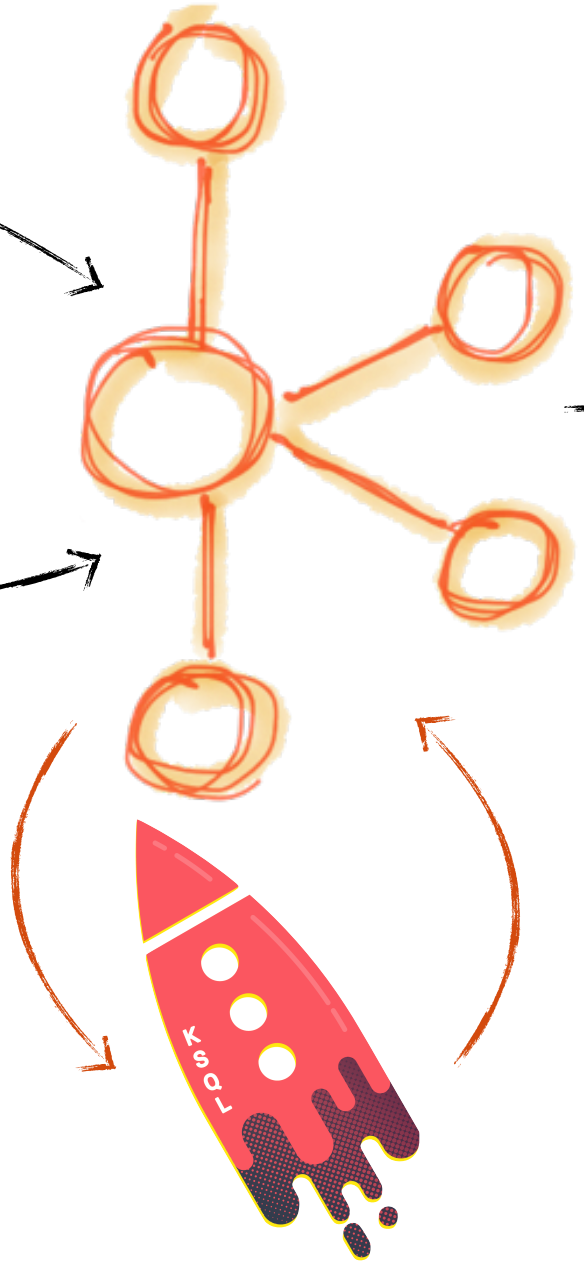
Debezium

Kafka Connect

Kafka Connect



Elasticsearch



```
ksql>
ksql>
```

```
asgard02  Screen(s): 0 bash  1 Elasticsearch  2 Kibana  3 mysql  4* ksql  5 datagen  16:35  Wed, Apr 18 2018  LD:3.14
```

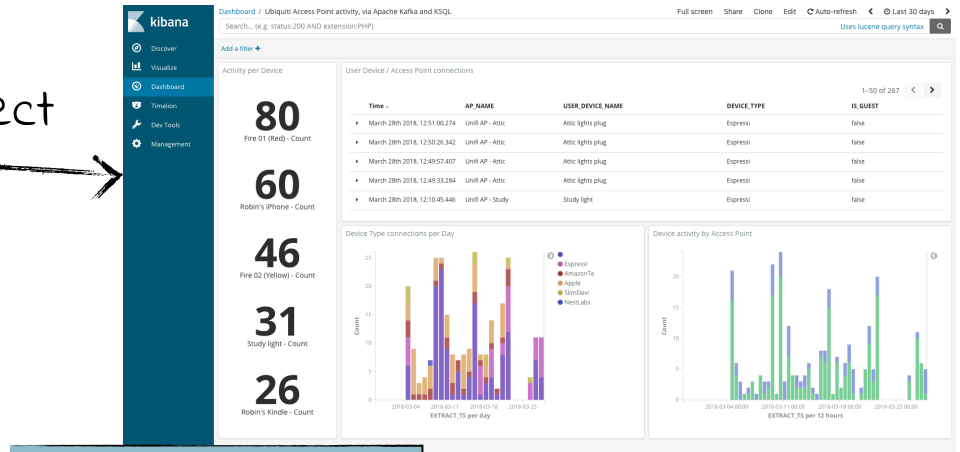
ratings stream

MySQL

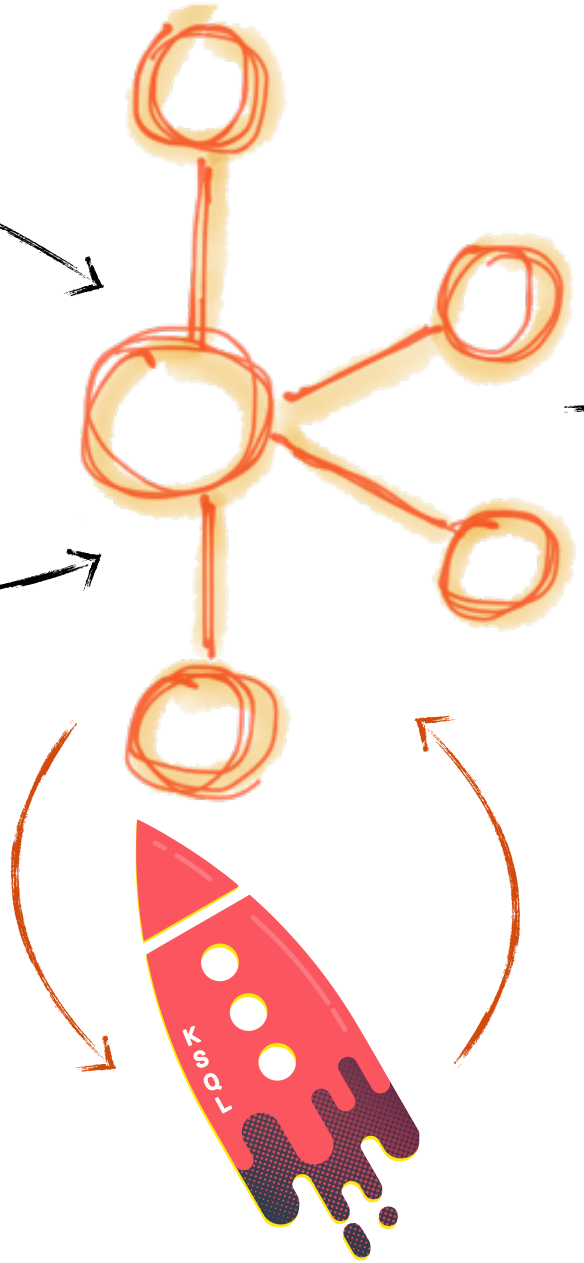
Debezium

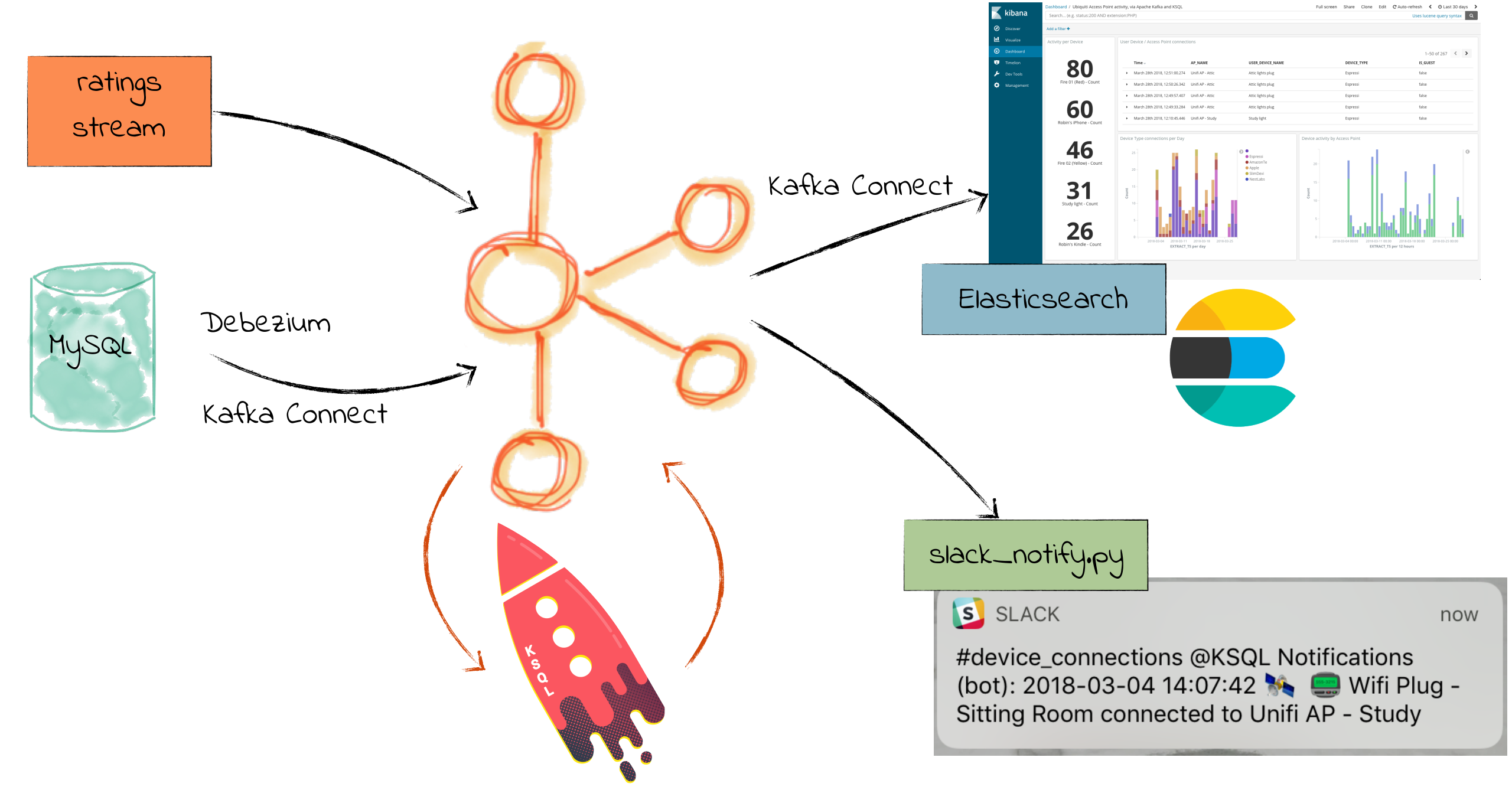
Kafka Connect

Kafka Connect

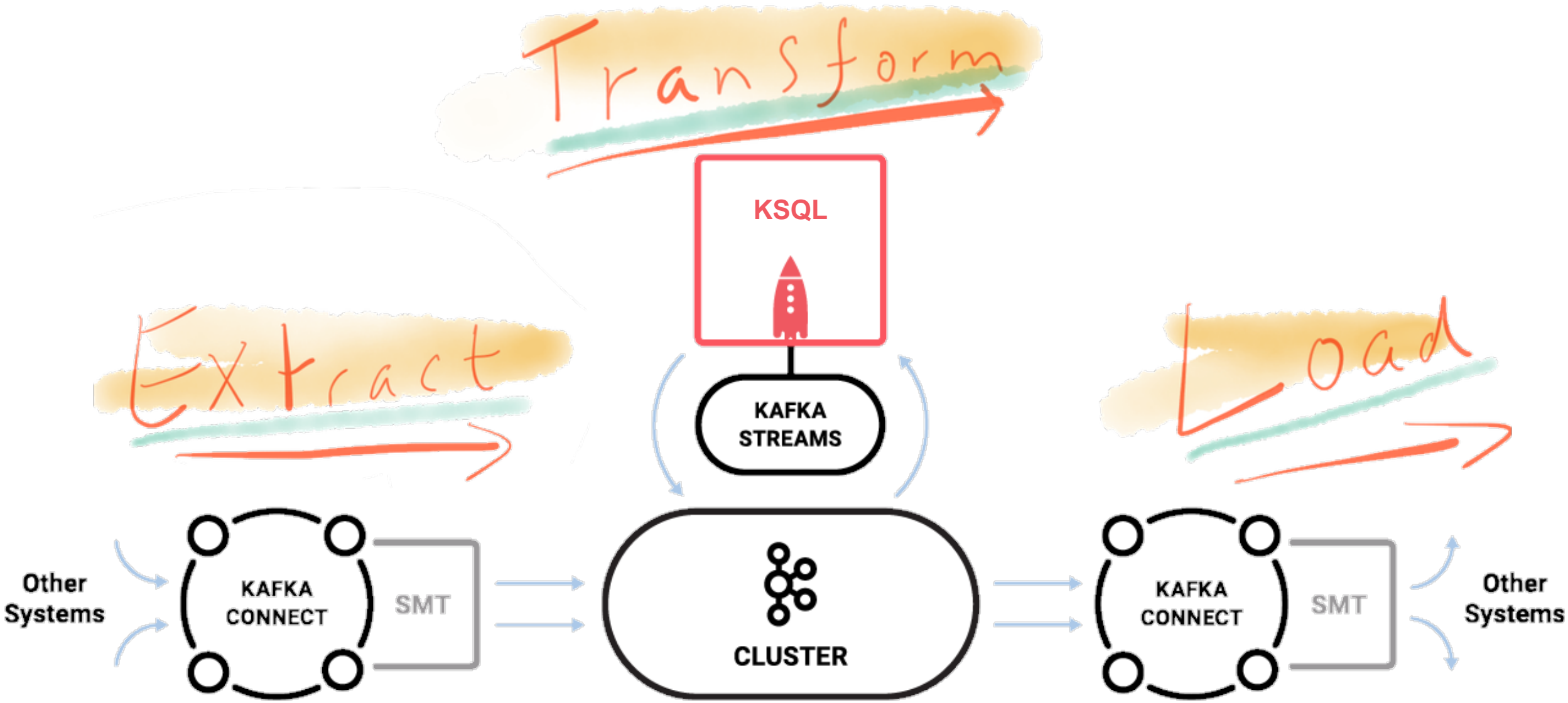


Elasticsearch

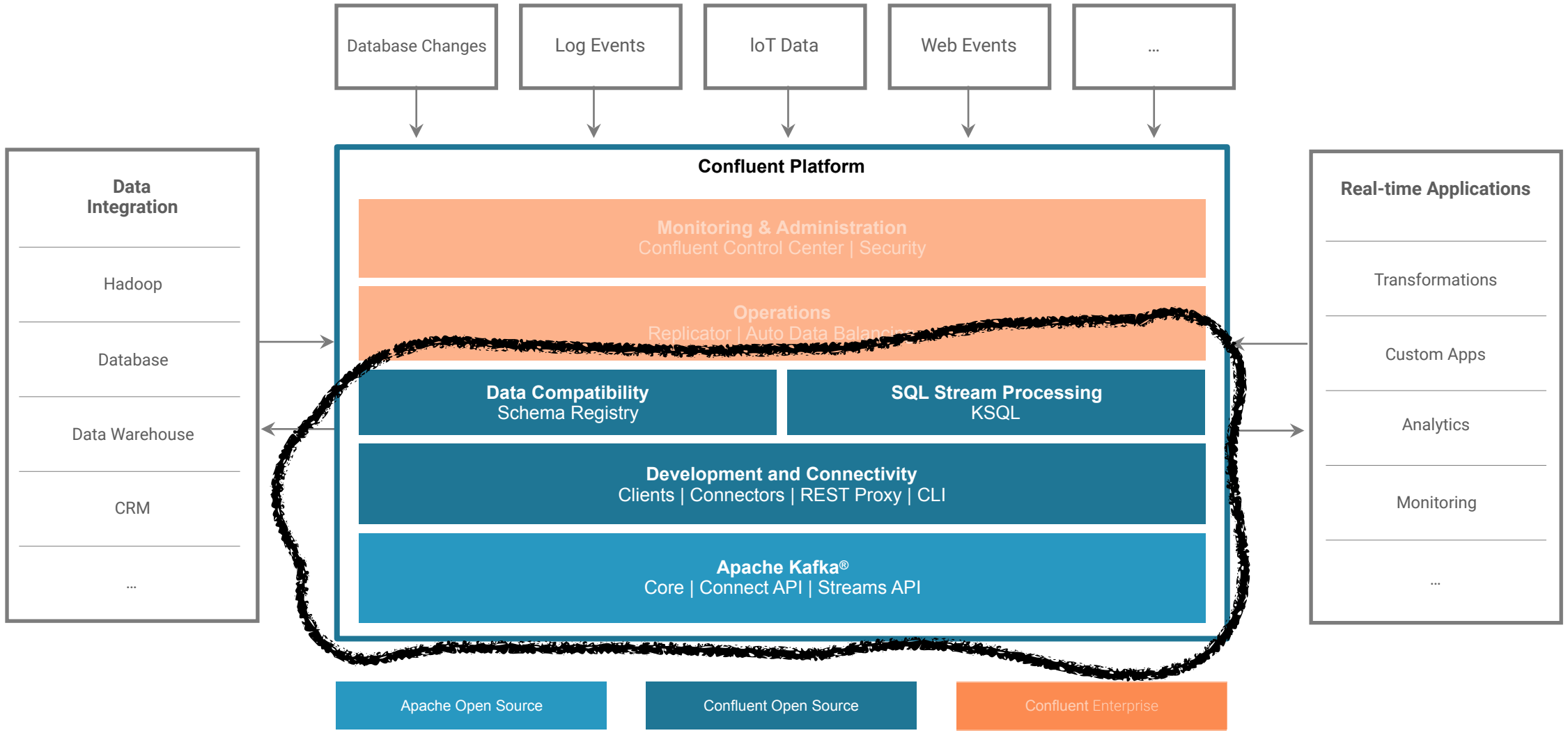


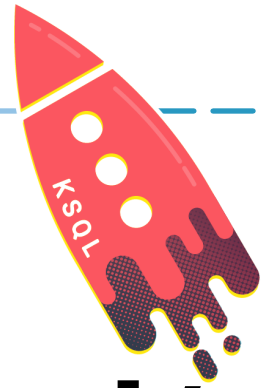

ratings stream
MySQL
Debezium
Kafka Connect
Kafka Connect
Elasticsearch
slack_notify.py
KSQL
SLACK
now
#device_connections @KSQL Notifications (bot): 2018-03-04 14:07:42 Wifi Plug - Sitting Room connected to Unifi AP - Study

Streaming ETL, powered by Apache Kafka and Confluent Platform



Confluent Open Source : Apache Kafka with a bunch of cool stuff! For free!





https://www.confluent.io/download/

https://slackpass.io/confluentcommunity

@rmoff robin@confluent.io

kafka summit
Kafka Summit San Francisco 2018
Dates: October 16-17, 2018
Location: Pier 27, San Francisco
CFP Opens: April 30, 2018
www.kafka-summit.org
Presented by confluent

https://speakerdeck.com/rmoff/

Look Ma, no Code! Building Streaming Data Pipelines with Apache Kafka

Kafka Summit London

24 Apr 2018 / Robin Moffatt

@rmoff robin@confluent.io